Taxon incompleteness and discrete time bins affect character change rates in simulated data

Flores, Jorge R.

2020-11-25

Flores , J R 2020 , ' Taxon incompleteness and discrete time bins affect character change rates in simulated data ' , Biology Letters , vol. 16 , no. 11 , 20200418 . https://doi.org/10.1098/rsbl.2020.0418

http://hdl.handle.net/10138/325704
https://doi.org/10.1098/rsbl.2020.0418

acceptedVersion

BIOLOGY LETTERS

royalsocietypublishing.org/journal/rsbl

Research

Cite this article: Flores JR. 2020 Taxon incompleteness and discrete time bins affect character change rates in simulated data. *Biol. Lett.* **16**: 20200418.
http://dx.doi.org/10.1098/rsbl.2020.0418

Received: 3 June 2020
Accepted: 19 October 2020

Subject Areas:
taxonomy and systematics, palaeontology, evolution

Keywords:
macroevolution, morphology, phylogenetics, taxon incompleteness, time bins

Author for correspondence:
Jorge R. Flores
e-mail: jorge.flores@helsinki.fi

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.5189343.

THE ROYAL SOCIETY PUBLISHING

Evolutionary biology

Taxon incompleteness and discrete time bins affect character change rates in simulated data

Jorge R. Flores

Finnish Museum of Natural History (Botany), University of Helsinki, PO Box 7, FI-00014, Helsinki, Finland

JRF, 0000-0002-2657-6126

Estimating how fast or slow morphology evolves through time (phenotypic change rate, PR) has become common in macroevolutionary studies and has been important for clarifying key evolutionary events. However, the inclusion of incompletely scored taxa (e.g. fossils) and variable lengths of discrete arbitrary time bins could affect PR estimates and potentially mask real PR patterns. Here, the impact of taxon incompleteness (unscored data) on PR estimates is assessed in simulated data. Three different time bin series were likewise evaluated: bins evenly spanning the tree length (i), a shorter middle bin and longer first and third bins (ii), and a longer middle bin and shorter first and third bins (iii). The results indicate that PR values decrease as taxon incompleteness increases. Statistically significant PR values, and the dispersion among PR values, depended on the time bins. These outcomes imply that taxon incompleteness can undermine our capacity to infer morphology evolutionary dynamics and that these estimates are also influenced by our choice of discrete time bins. More importantly, the present results stress the need for a better approach to deal with taxon incompleteness and arbitrary discrete time bins.

1. Introduction

Along with unravelling phylogenetic patterns, another goal of comparative and evolutionary biology is estimating how slow or fast character systems have evolved in different lineages through time [1–5]. Assessing the morphology change rate (phenotypic change rate, PR) improves our understanding of evolutionary events and helps to explain morphological diversity. PR estimates, for instance, were found to be coupled with morphological disparity and phyletic diversity in certain taxonomic groups [6,7]. Furthermore, in both extant and extinct groups, PRs were shown to be linked with events of adaptive radiations and climatic changes [8,9]. The complex nature of morphological data, however, imposes challenges to infer PR in some organisms [10].

Incompletely scored taxa characterize many morphological datasets and include two types of unscored entries: missing data ('?') and inapplicable characters ('–'). Both can be high in matrices that consist of fragmentary fossils [11] and can reduce taxon stability during phylogenetic inference (e.g. [12]). Some methods to estimate PR treat both unscored data types as 'NA' [4], while their treatment in others approaches—designed to infer divergence time based on molecular data—is rather unclear [13,14]. In Claddis [4], an R package, PRs are derived from the number of character state changes occurring along a branch of a certain length. Subsequently, PR values are deemed statistically high or low relative to other data partitions (e.g. PR values among time bins) based on a likelihood ratio test or an Akaike information criterion test. Although branch incompleteness is considered in estimates of PR, there is still no satisfactory approach to deal with missing data or inapplicable characters, and their

inclusion can lead to false negatives (i.e. not recognizing PRs statistically lower or higher than the remaining data partitions [4]). In empirical studies, the treatment of unscored data (either characters or taxa) has been puzzling. While some authors exclude characters or taxa without apparent reason other than being highly incomplete (e.g. [6]), others argue that incompleteness accentuates PR estimates [15]. Although taxon incompleteness is known to influence PR estimates [16], its precise impact is still uncertain since no explicit evaluations have been conducted [4,15–17].

Studies of evolutionary dynamics often involve discrete time bins (e.g. [6]), whose definition usually affects the inferred patterns [18–20]. In diversity studies, discretizing time bins leads to richness curves that depart from the expected 'correct' patterns in simulated data [20]. Likewise, discrete time bins have been deemed problematic in disparity-through-time (DTT) assessments [18]. In PR, however, the effect of varying time bin lengths, and its interplay with taxon incompleteness, have rarely been explored explicitly in empirical studies (e.g. [8,10,21]). Moreover, given that branches might span through multiple time bins, accounting for the influence of taxon incompleteness in different time bins is not straightforward [10]. Thus, the PR for a given time bin depends not only on branch completeness but also on branch lengths and time bin lengths. Altogether, the impact of both taxon incompleteness and variable time bin lengths on PR estimates remains unclear. In this study, therefore, the effect of different taxon incompleteness levels and time bin lengths on inferred PR values is evaluated in datasets simulated using equiprobable model trees and with missing data distributed both at random and according to empirical datasets. The outcomes shed light on the possible restrictions posed by taxon incompleteness in our understanding of morphological evolution.

2. Methodology

Datasets consisting of 20 taxa and 100 binary characters were simulated in TNT 1.5 [22] under a two-states Jukes–Cantor (JC) model (or 'Cavender–Farris–Neyman' model; [23–25]) wherein character states were equiprobably distributed among matrix cells. Datasets were simulated using equiprobable, ultrametric trees as a model with branch lengths randomly varying from 0.1 to 100. In these matrices, scored cells were replaced by missing data ('?') following two different approaches: 'random' and 'empirically based'. Both approaches differ on the distribution of missing entries: while the 'random' approach distributes missing entries equiprobably throughout the matrix, missing entries are distributed more frequently in certain regions of the matrix under the 'empirically based' protocol (see electronic supplementary material, figure S1). While the former approach has been widely used in previous studies (e.g. [11]), the reason for using the latter approach is that it reflects the different degrees of incompleteness observed among subsets of characters owing to their dissimilar preservation potential (e.g. [20]) (see electronic supplementary material).

In both cases, 100 initial completely scored matrices were subjected to four iterations where an increasing amount of scored cells were replaced with missing entries, achieving a maximum that oscillated between 15 and 50% of missing entries per dataset (which fit the values observed in empirical matrices; electronic supplementary material, table S1). Thus, the procedures generated 100 four-matrix series—each series simulated upon a certain model tree—wherein the only factor that varied was the amount of unscored data. Although larger matrices (e.g. >500 characters and >20 taxa) would increase the power of PR estimations, smaller datasets reduce the computational effort. The mid-sized matrices analysed here represented a compromise between both efficient computational effort and statistical power. Additionally, matrices with high character/taxon ratios—as in the empirical and simulated matrices considered herein—are more robust to variable analytical conditions as compared with smaller ones [26]. For an extended explanation on the methodology, see the electronic supplementary material.

PR was assessed in the simulated matrices by using three time bins and using the function 'DiscreteCharacterRate' of the R package Claddis [4,27]. PRs were estimated upon the model trees and calculated as the average number of character changes per branch length and branch completeness [4]. To evaluate the effect of using different bin lengths, three sets of (three) time bins were evaluated (i–iii) wherein bins spanned different proportions of the total tree length. Firstly, time bins were each set to span 0.3 of the entire tree length (i.e. time bins of equal length; i). Secondly, a shorter middle bin spanned 5% (from 0.3 to 0.25) and two bins covered 47.5% each of the total tree length (ii). Thirdly, a longer middle bin extended 70% (from 0.85 to 0.15) and two bins spanned 15% each of the tree length (iii). These three time bin sets were chosen to evaluate the effect of widely different bin lengths. Although these bin sets may represent exceptional cases, it should be noted that both extremely 'short' and 'long' time bins have been employed in empirical studies (e.g. [6]). To take into account branch incompleteness per time bin, the 'Lloyd' option from the 'DiscreteCharacterRate' function was employed (see details in Lloyd [4] and Claddis documentation). PR statistically high or low values within time bins—relative to the PRs estimates in the remaining bins—were evaluated through a likelihood ratio test (see electronic supplementary material). Finally, PR values per bin were plotted against the proportion of missing data for the three time bin series considered (i–iii).

3. Results

The analyses showed that, for these simulated datasets, PR values commonly diminished as the proportion of missing data increased: leading to similar patterns under both the 'random' and 'empirically based' approaches (figure 1; electronic supplementary material, figure S2). Only in a single time bin (shorter middle time bin) under the 'random' approach, missing data had nearly no impact on PR (electronic supplementary material, figure S2). In these JC simulated datasets, PR values in the first time bin tended to be statistically higher than in the remaining bins (red dots, figure 1) while statistically lower PR values were more frequent in the second and third time bins (blue dots, figure 1). Given that both approaches led to similar patterns, only the results of the 'empirically based' protocol will be discussed.

Varying the extension of time bins affected the PR estimates (*a*–*c*, figure 1; electronic supplementary material, figure S2). As compared with even time bins (*a*), uneven time bin series (*b*,*c*) altered observed PR values. More specifically, shortening time bins increased the dispersion and reduced the proportion of statistically significant PR values while the opposite trend was seen in longer time bins (figures 1 and 2; electronic supplementary material, figure S2). For instance, when the middle time bin was shorter (*b*), the dispersion among PR values increased, and the proportion of significant PR values was minimized as compared with the other middle time bins (figures 1 and 2; electronic supplementary material, figure S2). This outcome

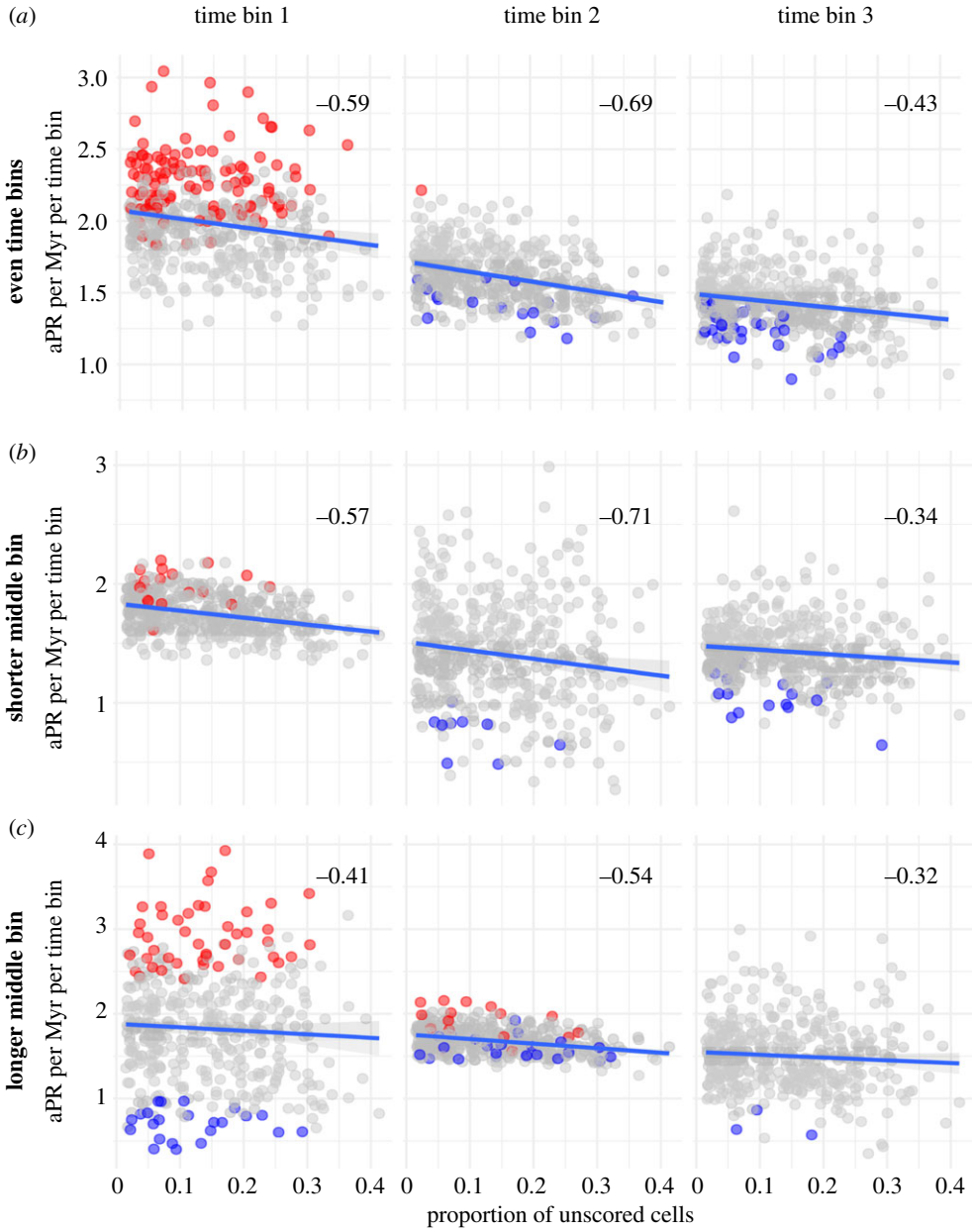

Figure 1. Average number of character state changes (average phenotypic change rate, aPR) per Myr in each time bin against the percentage of unscored cells, under the 'empirically based' approach and the different time bin series (*a–c*). Red and blue symbols indicate statistically high and low values, respectively. Grey represents no statistical significance. Slope of tendency line is indicated in each plot.

was reversed in longer middle time bins (*c*, figures 1 and 2; electronic supplementary material, figure S2). Note that PR values within the other time bins were also affected complementarily after modifying their extension (figure 1; electronic supplementary material, figure S2). In fully scored matrices, a similar response was seen in terms of statistically significant values after modifying time bin lengths: shorter time bins reduced the proportion of either high or low values (figure 2). Thus, adding unscored data enhanced the effect caused by different bin lengths.

4. Discussion and final remarks

The present estimations of the number of character change relative to branch lengths (phenotypic change rate, PR) were sensitive to both taxon incompleteness and the time bins selected (figure 1). Taxon incompleteness has long been studied in the context of phylogenetic inference (e.g. [28]), though its impact on PR estimation is not fully certain. While some authors have excluded characters or taxa that were highly incomplete [29,30], taxon incompleteness has not been explicitly addressed in other studies (e.g. [7,14]). In early tetrapods, the inclusion of fragmentary fossils was argued to exaggerate both fast and slow PRs [15]. Following this logic, the number of statistically high or low PR values should increase as more unscored data are sampled. However, in the present analyses, the response of PR values to the proportion of unscored data was often negative (figure 1). For instance, if incompleteness accentuated PR [15], PR values in the first time bin should have increased as more unscored cells were present; instead, the opposite pattern is seen here (figure 1). The decay in PR observed here as a consequence of unscored data was also seen in DTT metrics [31]. Overall, this result expands upon previous analyses that showed the negative impact of incompleteness on disparity measures [31], and indicate that the inclusion of highly fragmentary taxa could hinder the distinction of an otherwise high phenotypic rate.

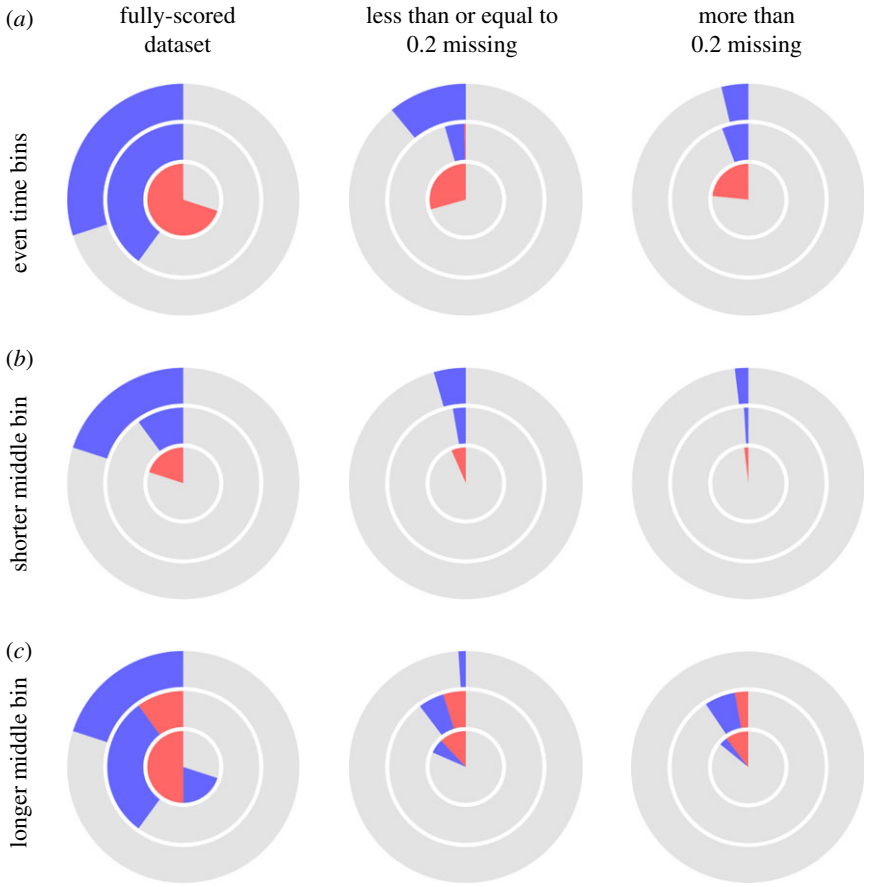

Figure 2. Effect of varying time bin lengths on the proportion of statistically significant PR values: low (blue) or high (red). Time bins are represented as rings in the bullseye: first time bin (innermost ring), middle time bin (second ring) and third time bin (external ring). Proportions are counted for the fully scored datasets, datasets with less than or equal to 0.2 missing entries and datasets with more than 0.2 missing data.

Another interesting outcome is the effect of varying the limits of the time bins on the statistical significance of the PR values and the dispersion of PRs within time bins: shorter bins reduced the number of statistically significant values and increased dispersion (figures 1 and 2). In fully scored matrices, differential time bin lengths also modified PR estimates, a pattern enhanced by missing data (figure 2). As already mentioned, addressing branch incompleteness is not straightforward given that branches commonly span time bin limits (see [4] and Claddis R package documentation). In DTT studies, it has been already noticed that the usage of discrete time bins can introduce biases [18,31]. Time bins are commonly based on stratigraphic data and, therefore, are of unequal lengths (e.g. [6,15,32]). In those cases, biases are due to the uneven number of taxa included in different time bins: disparity is higher in longer time bins wherein more taxa are included [18,31]. Although rarely discussed, bin lengths can also affect the statistical power of the PR test. This is because more branches are sampled in longer time bins. Therefore, as the sampling size increases, the statistical power of the test improves. Conversely, when time bins are shorter, fewer branches are sampled and the statistical power is reduced. In the present analyses, the proportion of significant rates within time bins was reduced after shortening the bin limits, and this is exacerbated by increasing proportions of missing data (figure 2). It is worth noting that the number of branches present in a specific bin also depends on the topology of the tree: symmetric trees are likely to have more branches included in a time bin as compared with asymmetric trees.

In actual matrices, taxon incompleteness involves both missing data and inapplicable character states. Although the two data types are different in nature, they are commonly treated equally in current implementations to evaluate PR (e.g. [4]). To take into account incompleteness, both weighting ('Close', [4,10]) and 'subtree' ('Lloyd', see details in [4]) approaches were proposed. Another proposal involves down-weighting characters that 'concentrate' a large number of the observed changes [4], in a similar way to implied weighting [33]. Nevertheless, since missing data ('?') and inapplicable characters ('–') are treated equally, these approaches can overestimate incompleteness. A potential way to discriminate both data types is based on the usage of step-matrix characters. If considering inapplicable character states as additional states, prohibitive transformation costs between these and scored character states could be employed to avoid treating inapplicable data as missing (NA). Such an approach requires no ad hoc measures (e.g. weighted means, [10]) since inapplicable character state changes are inferred only between terminal nodes and their last ancestor node. Branch incompleteness is thus due to missing data or ambiguous character state reconstruction. However, because step-matrix characters are not yet supported in Claddis, implementing this strategy is not possible.

To conclude, the present study shows that missing data and the choice of different time bin lengths affect PR estimates. These outcomes stress the need for approaches that could deal with unscored data and avoid biases emerging from the definition of discrete time bins. The completeness of the fossil record varies through geological time and is dependent on the taxonomic group [34–36]. Therefore, even though the

amount of scored data could be improved as more complete fossils are discovered, this will be hardly achieved for some taxa. Moreover, since the definition of discrete time bins affects PR—and other metrics as well [18,19]—methodological improvements are required to deal with incompleteness in the context of macroevolutionary studies. In particular, performing sensitivity analyses (varying time bin lengths) or devising methods for using continuous time bins [18] will prove beneficial for PR studies.

Data accessibility. Scripts to simulate data and conduct character change rate inference are provided as electronic supplementary material.

Competing interests. I declare I have no competing interests

Funding. This study was supported by FONCyT (grant no. PICT 0810) and by PIUNT (grant no. G631).

Acknowledgements. The author thanks Christopher Dean, Edward Braun and an anonymous reviewer for their constructive comments on an early version of the manuscript. Graeme Lloyd kindly commented on a previous pre-print version of this manuscript. TNT is free to users thanks to the Willi Hennig Society.

References

1. Schaeffer J, Benton MJ, Rayfield EJ, Stubbs TL. 2020 Morphological disparity in theropod jaws: comparing discrete characters and geometric morphometrics. *Palaeontology* **63**, 283–299. (doi:10.1111/pala.12455)
2. Benton MJ. 1995 Available diversification and extinction in the history of life. *Science* **268**, 52–58. (doi:10.1126/science.7701342)
3. Li Q, Ni X. 2016 An early Oligocene fossil demonstrates treeshrews are slowly evolving 'living fossils'. *Scient. Rep.* **6**, 18627. (doi:10.1038/srep18627)
4. Lloyd GT. 2016 Estimating morphological diversity and tempo with discrete character–taxon matrices: implementation, challenges, progress, and future directions. *Biol. J. Linn. Soc.* **118**, 131–151. (doi:10.1111/bij.12746)
5. Benton MJ, Donoghue PCJ. 2007 Paleontological evidence to date the tree of life. *Mol. Biol. Evol.* **24**, 26–53. (doi:10.1093/molbev/msl150)
6. Ezcurra MD, Butler RJ. 2018 The rise of the ruling reptiles and ecosystem recovery from the Permo-Triassic mass extinction. *Proc. R. Soc. B* **285**, 20180361. (doi:10.1098/rspb.2018.0361)
7. Brocklehurst N. 2017 Rates of morphological evolution in Captorhinidae: an adaptive radiation of Permian herbivores. *PeerJ.* **5**, e3200. (doi:10.7717/peerj.3200)
8. Slater GJ, Price SA, Santini F, Alfaro ME. 2010 Diversity versus disparity and the radiation of modern cetaceans. *Proc. R. Soc. B* **277**, 3097–3104. (doi:10.1098/rspb.2010.0408)
9. Marx FG, Fordyce RE. 2015 Baleen boom and bust: a synthesis of mysticete phylogeny, diversity and disparity. *R. Soc. Open Sci.* **2**, 140434. (doi:10.1098/rsos.140434)
10. Close RA, Friedman M, Lloyd GT, Benson RBJ. 2015 Evidence for a mid-Jurassic adaptive radiation in mammals. *Curr. Biol.* **25**, 2137–2142. (doi:10.1016/j.cub.2015.06.047)
11. Wiens J. 2006 Incomplete taxa, incomplete characters, and phylogenetic accuracy: is there a missing data problem? *J. Vertebr. Paleontol.* **23**, 297–310. (doi:10.1671/0272-4634(2003)023[0297:iticap]2.0.co;2)
12. Pol D, Escapa IH. 2009 Unstable taxa in cladistic analysis: identification and the assessment of relevant characters. *Cladistics* **25**, 515–527. (doi:10.1111/j.1096-0031.2009.00258.x)
13. Bouckaert R *et al.* 2019 BEAST 2.5: an advanced software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **15**, e1006650. (doi:10.1371/journal.pcbi.1006650)
14. Varela L, Tambusso PS, Mcdonald HG, Fariña RA. 2019 Phylogeny, macroevolutionary trends and historical biogeography of sloths: insights from a Bayesian morphological clock analysis. *Syst. Biol.* **68**, 204–218. (doi:10.1093/sysbio/syy058)
15. Ruta M, Wagner PJ, Coates MI. 2006 Evolutionary patterns in early tetrapods. I. Rapid initial diversification followed by decrease in rates of character change. *Proc. R. Soc. B* **273**, 2107–2111. (doi:10.1098/rspb.2006.3577)
16. Brusatte SL, Lloyd GT, Wang SC, Norell MA. 2014 Gradual assembly of avian body plan culminated in rapid rates of evolution across the dinosaur-bird transition. *Curr. Biol.* **24**, 2386–2392. (doi:10.1016/j.cub.2014.08.034)
17. Lloyd GT, Wang SC, Brusatte SL. 2012 Identifying heterogeneity in rates of morphological evolution: discrete character change in the evolution of lungfish (Sarcopterygii; Dipnoi). *Evolution* **66**, 330–348. (doi:10.5061/dryad.pg46f)
18. Guillerme T, Cooper N. 2018 Time for a rethink: time sub-sampling methods in disparity-through-time analyses. *Palaeontology* **61**, 481–493. (doi:10.1111/pala.12364)
19. Dean CD, Chiarenza AA, Maidment SCR. In press. Formation binning: a new method for increased temporal resolution in regional studies, applied to the Late Cretaceous dinosaur fossil record of North America. *Palaeontology*. (doi:10.1111/pala.12492)
20. Gibert C, Escarguel G. 2017 Evaluating the accuracy of biodiversity changes through geologic times: from simulation to solution. *Paleobiology* **43**, 667–692. (doi:10.1017/pab.2017.10)
21. Paterson JR, Edgecombe GD, Lee MSY. 2019 Trilobite evolutionary rates constrain the duration of the Cambrian explosion. *Proc. Natl Acad. Sci. USA* **116**, 4394–4399. (doi:10.1073/pnas.1819366116)
22. Goloboff P, Catalano S. 2016 TNT version 1.5, including a full implementation of phylogenetic morphometrics. *Cladistics* **32**, 221–238. (doi:10.1111/cla.12160)
23. Cavender JA. 1978 Taxonomy with confidence. *Math. Biosci.* **40**, 271–280. (doi:10.1016/0025-5564(78)90089-5)
24. Farris JS. 1973 Probability model for inferring evolutionary trees. *Syst. Biol.* **22**, 250–256. (doi:10.1093/sysbio/22.3.250)
25. Neyman J. 1971 Molecular studies of evolution: a source of novel statistical problems. In *Statistical decision theory and related topics* (eds SS Gupta, J Yackel), pp. 1–27. New York, NY: New York Academic Press.
26. Bremer B, Jansen R, Oxelman B, Backlund M, Lantz H, Kim K. 1999 More characters or more taxa for a robust phylogeny—case study from the coffee family (Rubiaceae). *Syst. Biol.* **48**, 413–435. (doi:10.1080/106351599260085)
27. R Core Team. 2019. *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. See https://www.r-project.org/.
28. Wiens J, Morrill MC. 2011 Missing data in phylogenetic analysis: reconciling results from simulations and empirical data. *Syst. Biol.* **60**, 719–731. (doi:10.1093/sysbio/syr025)
29. Sidor CA, Hopson JA. 1998 Ghost lineages and 'mammalness': assessing the temporal pattern of character acquisition in the Synapsida. *Paleobiology* **24**, 254–273. (doi:10.1666/0094-8373-24.2.254)
30. Rowe T. 1988 Definition, diagnosis, and origin of Mammalia. *J. Vertebr. Paleontol.* **8**, 241–264. (doi:10.1080/02724634.1988.10011708)
31. Smith AJ, Rosario MV, Eiting TP, Dumont ER. 2014 Joined at the hip: linked characters and the problem of missing data in studies of disparity. *Evolution* **68**, 2386–2400. (doi:10.1111/evo.12435)
32. Brusatte SL, Montanari S, Yi H, Norell MA. 2011 Phylogenetic corrections for morphological disparity analysis: new methodology and case studies. *Paleobiology* **37**, 1. (doi:10.1666/09057.1)
33. Goloboff P. 1993 Estimating character weights during tree search. *Cladistics* **9**, 83–91. (doi:10.1111/j.1096-0031.1993.tb00209.x)
34. Verrière A, Brocklehurst N, Fröbisch J. 2016 Assessing the completeness of the fossil record: comparison of different methods applied to parareptilian tetrapods (Vertebrata: Sauropsida). *Paleobiology* **42**, 680–695. (doi:10.1017/pab.2016.26)
35. Davies TW, Bell MA, Goswami A, Halliday TJD. 2017 Completeness of the eutherian mammal fossil record and implications for reconstructing mammal evolution through the Cretaceous/Paleogene mass extinction. *Paleobiology* **43**, 521–536. (doi:10.5061/dryad.r0881)
36. Dean CD, Mannion PD, Butler RJ. 2016 Preservational bias controls the fossil record of pterosaurs. *Palaeontology* **59**, 225–247. (doi:10.1111/pala.12225)